Optimistic Active Learning using Mutual Information

Yuhong Guo and Russell Greiner
University of Alberta, Edmonton, Canada

**Idea:** an optimistic active learner that exploits the **discriminative partition information** in the unlabeled instances, makes an **optimistic assessment** of each candidate instance, and temporarily switches to a different policy if the optimistic assessment is wrong.

**Optimistic Query Selection**

1. Most uncertain query selection (MU):
   \[
   \arg \max_{i \in U} H(Y_i | x_i, L) = \sum_i P( y_i | x_i, L) \log P( y_i | x_i, L)
   \]
   **Shortcoming:** ignores the unlabeled data !

2. Select the query that maximizes its conditional mutual information about the unlabeled data:
   \[
   \arg \max_{i \in U} \{ H(Y_i | X_U, L) - H(Y_i | X_U, L, (x_i, y_i)) \}
   \]
   **Proposals:**
   (a) Take the expectation wrt \( Y_i \) (MCMI[avg]):
   \[
   \arg \min_{i \in U} \sum_i H(Y_i | x_i, \theta_{L+}(x_i, y_i))
   \]
   **Shortcoming:** aggravates the ambiguity caused by the limited labeled data.
   (b) Take an **optimistic** strategy: use only the best query label (MCMI[min]):
   \[
   \arg \min_{i \in U} \sum_i H(Y_i | x_i, \theta_{L+}(x_i,))
   \]
   **Question:** How to determine \( y_i \)?

   **Proposals:**

   - Take the expectation wrt \( Y_i \) (MCMI[avg]):
     \[
     \arg \min_{i \in U} \sum_i H(Y_i | x_i, \theta_{L+}(x_i,))
     \]
   - Online Adjustment:
     - Can easily detect this “guessed wrong” situation, in the immediate next step,
     - Simply compare the actual label for the query with its optimistically predicted label
   - Whenever Mm+M guesses wrong,
     - it switches to a different query selection criterion (MU) for the next 1 iteration

**Experimental Evaluation**

- **Comparing Mm+M with other Active Learners**
  - Over 100 sample sizes, over 17 datasets:
    - Mm+M was
      - “statistically better” 85 times
      - “statistically worse” 2 times
      - “tied” 13 times
    - Signed Rank Test shows Mm+M is better

- **Comparing Mm+M vs MU, for PIMA dataset:**
  - Over 100 sample sizes, Mm+M was
    - “statistically better” for >5 more sample-sizes: 13 times
    - “statistically worse” for >5 more sample-sizes: 2 times
    - Signed Rank Test shows Mm+M is better

- **Future work:**
  - Understand when Mm+M is appropriate
  - Design further variants